DATA MANAGEMENT IN THE HUMANITIES

David G. Hays

April 1968

P-3834

DATA MANAGEMENT IN THE HUMANITIES

David G. Hays

RAND Corporation
Santa Monica, California

April 1968

DATA MANAGEMENT IN THE HUMANITIES

David G. Hays*

The RAND Corporation, Santa Monica, California

Only man possesses speech, although most animal life has some limited means of communication.

Only man possesses values, although animal life is generally motivated by hunger, thirst, fear, and sexuality.

Only man possesses art, although the wings of a butterfly, the towering hills of exotic ants, and the courting patterns of a few birds and fish are beautiful to man.

The proper study of mankind is man, and the humanities encompass -- in their study of speech, values, and art -- what is most characteristically human. It would be obfuscatory, and a renunciation of the values which humanistic scholarship is devoted to examining, to subsume humanistic scholarship under the rubric of science; this would be scientism of the lowest order. Science does, however, proceed in accordance with its own pattern of values, and to study them is an appropriate activity for

---

*Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff.

This paper was prepared for presentation at the Fifteenth Annual International Technical Communications Conference of the Society of Technical Writers and Publishers, May 10, 1968, Los Angeles, California.

humanists. Science has form, and to evaluate the form of science in accordance with aesthetic criteria is appropriate also.

But the humanities are more strongly distinguished from science by their ultimate criteria of value than by their techniques. An aesthetic valuation of a painting that leaves out of account the chemistry of pigments and of retinal response, or one that, through ignorance, makes false assumptions about these chemistries, is asinine. Early Gestalt psychology is not as good a basis for the study of architectural form as the modern psychological theory of pattern recognition and cognition in the human species. The analyst of style is not as well served by the older syntax and semantics of high-school grammar and lexicography as we hope he will be served by the new syntax and semantics of mathematical linguistics. Even the cabinet maker must know his material; the scholar cannot escape this necessity, and it is science which tells him what the stuff of art is, leaving to him the question of how it is.

Let us distinguish between art, the creation of beauty; and humanistic scholarship, the comprehension of beauty. For the humanities, the data are primarily objects of art -- pictures, buildings, texts, dances -- and human responses to those objects. Like the rest of man's intellectual endeavors, humanistic scholarship has second-

ary data -- the analytic responses of past scholars to the same primary data.

Human life began when speech began. Scholarship became possible with the invention of writing, the externalization of the human ability to store information. Simulation of human activity became possible, we believe, with the invention of the computer and of the mathematics that makes computation possible. We have scarcely begun to examine the most salient features of the problem that simulation of human capacities will inevicably raise. How simple it would be if man could create machines in his own image and abdicate. Or better, man might deny the machine free will, whether man possesses it or not, and make himself god, the machines his parishioners. John Barth, in The Sotweed Factor and Giles Goat Boy has shown us a few of the results that obtain if lackey or goatherd turns god. In common with the rest of life, man has always had the ability to reproduce himself, and the grossly excessive population of this planet has not caused some of us to lose respect for the dignity of every single human life. The possibility of creating man-like computers, programs that can simulate the most human aspects of man's behavior, need not cause us to lose our sense of the unity of mankind, or our xenophobic attitude toward the machine. Nor need it make us respect the computer programmer more highly than the human scientist,

poet, or statesman.

The ability to simulate a process by no means implies understanding of the process. Let us imagine a computer program that writes sonnets. Do we understand the sonnet completely if we enjoy the sonnets written by the computer? Indeed we do not, and a long, arduous, and possibly even fruitless program of research must follow if the automatic construction of sonnets is to lead to aesthetic understanding of the sonnet as verse. The thrill of creating a simulacrum of man is not to be confused with the satisfaction of understanding him.

The work of scholarship remains, therefore, as it always has been, the increase of understanding by analytic study. Because the human response to beauty is universal, and refinable by education, the comprehension of beauty is a fitting goal for the general education of every human being, from kindergarten through the end of formal schooling and beyond. Because both the primary and secondary classes of data suitable for humanistic study and education are vast, and because the value of an individual work, like the value of an individual human being, is not depressed by the simultaneous existence of any number of similar works or persons, no matter how many, the scholar and the student need data management facilities of the finest sort.

The global community of scholars communicates more readily, thanks to television, radio, jet aircraft, communication satellites, and direct dial telephone systems. Selected materials are more widely available, thanks to the mass printing and mass circulation of paperback books. Motion pictures and theatres and schools contribute to the flow of primary and secondary data. J.C.R. Licklider has anticipated in Libraries of the Future in some measure the improvement that may be expected with world-wide access to a single body of data (and programs) stored in a computer and coupled to scholars' typewriters.

Perhaps the majority of humanistic scholars devote themselves to the study of texts. The library is the great traditional administrative device for putting texts at the disposal of scholars, and must remain so for a time. Librarians are less than ever before keepers of books; they are coming to be managers of data. I wish to speak of two techniques that will, in my opinion, serve them well. One uses computers, the other uses microphotography.

For about ten years, my associates and I at The RAND Corporation have been concerned with the availability of texts for study with the aid of computers. Martin Kay and Theodore W. Ziehe devised a pair of schemes which provide adequate means for recording text on magnetic tape. Kay and Ziehe pointed out that a standard was as necessary for the growth of libraries of digitally recorded text as

for the growth of the long-playing record business. The chosen width and depth of groove, and speed of rotation, are not markedly superior to very similar choices, but unless a large collection of discs are available with common critical dimensions, no one can afford to buy the instruments for playing them back.

The situation in computing is somewhat special. In different circumstances, different input and output devices are used: sometimes a paper tape typewriter, sometimes a special typesetter's keyboard, sometimes a console directly connected to a computer. All these and many other machines are good for recording text digitally. Line printers, typewriters, and photo-composition machines are good ways of converting digitally recorded text into graphic form, depending on the purpose in view. No single encoding scheme is appropriate to all these input and output devices; the computer itself solves the problem, because it can convert a standard, archival code into control signals for any printer, and it can follow the motions of any input typewriter, coding not the motions themselves, but the text signified by those motions, for permanent retention and interchange. These themes were opened by Kay and Ziehe in Natural Language in Computer Form, and developed in several papers from time to time.

Text has structure, but the structure of a text, like the structure of any other natural object, depends

on your point of view. The main structure of a text, for the editor, consists of chapters, sections, paragraphs, and sentences. For the typographer, it consists of pages and lines. The Kay-Ziehe scheme for text on tape is merely a way of imposing structure, whatever interpretation may be given it. It is a good way of recording bibliographies, the texts of books, or personnel files. This scheme, also, was first presented in Natural Language in Computer Form.

Once it is recognized that a generalized scheme of structural recording can be handled by computers without regard to the specific organization of a given text, and that different files can be given more or less structure, according to purpose; once it has been recognized that a text encoding scheme must allow for an indefinitely large character set, so that the ultimate richness of the printer's system can be recorded, but that a simple typescript can be recorded in such a system without undue difficulty -- then, and the time has come, libraries of text on magnetic tape can begin to grow in exactly the same way libraries of printed books have grown.

For the university of today and tomorrow, for the publisher, for the research institute, the library of digitally recorded text is a necessary component. It should not be a part of the computer center, any more than the library should be a part of the university press --

or, by another analogy, the corporate library a part of the typing pool. Most of the text in any institutional library will be obtained by purchase, in digitally recorded form, just as most of the books in a library are obtained by purchase.

Libraries of digitally recorded text will be used only for special purposes, at least for a while. These special purposes are analyses: scientific and humanistic studies of the construction of texts, in accordance with the existential criteria of science and the aesthetic criteria of the humanities. Students and teachers will consult the digital library when their studies go to the level of minute detail. Publishers, of course, will use their libraries for preparation of new editions, or recombinations, of older works.

Let it be noted in passing that the most natural suppliers of digitally recorded text are the publishers of ordinary books. The price of a book in digital recording will naturally be higher than the price on paper, but the tape will be a by-product of computerized typesetting, and therefore not as expensive as one might anticipate. I know of no publisher who includes in his current catalog a price for this form of work.

Now let us turn to a second method of improving the librarian's ability to manage data in the humanities. This is the method of mass-produced, high-reduction

microform libraries. RAND has just released a report called A Billion Books for Education in America and the World -- A Proposal. For me, this idea is more than four years old. Others have thought of it, some undoubtedly before I did. In Libraries and Universities, Paul Buck wrote of the value a master file of a million volumes on microform would have -- he was thinking of the needs of new campuses in the United States. William R. Polk, Director of the Stevenson Institute in Chicago, felt that large microform libraries, produced in identical sets, would be advantageous in providing for the needs of developing countries.

Here is our version of the plan: Using the massed skills of a great many experts in different fields, a collection of about a million books would be established. But not every item in the library need be a book -- there could be runs of journals, newspapers, pamphlets, and other scattered materials. There could be an important collection of manuscripts.

Once the titles had been selected, good copies would be sought out wherever they were and photographed on microfilm with the utmost care. Then the images would be transferred to a high-reduction form, in a final ratio of 60 to 200 diameters. As many as two or three thousand pages might be carried on one 4" x 6" transparency. An ultramicrofiche can be protected with transparent plastic,

and dirt or marring on the surface is of little consequence to the reader, since the depth of focus is so small that the surface of the plastic coating is invisible on the screen of a projector.

While the books were being photographed, they could be catalogued. Since the library would be catalogued only once, for the benefit of many institutions, the cataloging could be done with great care. Bibliographies could be extracted from the master catalog, and far more information about each item could be included in the catalog. The shelf list would serve as a browsing aid, since the library's customers would not be able to wander through the shelves, looking into book after book for matters of interest. The catalog could be published on microform or in printed books.

How large would a collection of a million volumes be, if manufactured in this form? With ten books per fiche, the library could be arranged on the surface of a counter fifty feet long, or a bit less; that would seem a convenient length for the circulation counter in a large, busy library. Or, if the material were two levels deep, the counter could be shortened by half, and so on.

How much would a collection cost? The lowest achievable price, if hundreds of sets could be sold, with no commercial risk, and with contributed labor for some of the intellectual tasks, would be rather less than two

hundred thousand dollars. Equipment would be a major additional expense, but the total is within the reach of every serious educational institution in the country -- but not within the present budget of every such institution.

Should we now cease printing books, and use libraries of ultramicrofiche and digitally recorded text for all purposes? Clearly not. My young children are enjoying the paperback revolution, which has just recently penetrated the elementary school. I hope that they can enjoy it for many years to come, and that their children too will take paperbacks to bed. I hope myself to be able to afford full calf bindings for part of my library, someday. Between these extremes there is an enormous field for the book, the magazine, and the scientific or scholarly journal as we know them. There is also a place for standard microform, ultramicroform, and digitally recorded text. The interesting problem in data management for the humanities is to achieve the right combination.

And what shall we do about copyright? If text is to be stored in disc packs, which look rather like nickelodeons, shall we just add a coin box? How can the low cost of ultramicroform be reconciled with the high cost of royalties? Is piracy about to destroy the legitimate publisher? It need not be so. In our stretch toward the future we are always hampered a little by the organiza-

tional supports we built for the technology of the past. A competitive publishing industry is the most important realization of free speech, and royalties are what stimulates competition. Unchecked piracy can only be followed by a reduction of freedom. We will enjoy the freedom of new technologies for data management if we succeed in building comfortable new organizational supports to fit them.